Visualising Topics in Document Collections

An Analysis of the Interpretation Processes of Historians

Anastasia Christoforidis, Ben Heuwing, Thomas Mandl

University of Hildesheim, Germany
anastasiachristoforidou@gmail.com, {heuwing, Mandl}@uni-hildesheim.de

Abstract

This paper discusses two multivariate visualisations which provide insights into topic model distributions across sub-collections of a collection of historical textbooks in the context of a digital humanities project. Results of a qualitative user study with experts in historical research indicate that network graphs are more appropriate for revealing general connections among sub-collections, while small-multiples of heatmaps of topic correlations are more suitable for a finer grained analysis of the connections between specific topics. We analyse the user behaviour during analysis to identify general activities of the interpretation of topic models as well as activities of interpreting visual elements that are specific to each visualisation. As a result, we observed usability problems and show potential for improved visualisations in digital humanities applications.

Keywords: information visualisations; topic modelling; digital humanities; evaluation; user study

1 Introduction

In the humanities, access to large collections of digitised sources has created an increased interest into the use of tools to support the analysis of the contents of large amounts of documents. While in disciplines such as literature

In: M. Gäde/V. Trkulja/V. Petras (Eds.): Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017), Berlin, 13th—15th March 2017. Glückstadt: Verlag Werner Hülsbusch, pp. 37—49.

studies, analysis takes into account mainly stylistic characteristics of documents, historians largely depend on an analysis at a topical level to make sense of a collection and to retrieve documents of interest. Modelling a collection and providing interactive visualisations of its contents are important and interdependent steps that are necessary to offer support for the research interests of researchers in the humanities, many of whom do not have a background in information technology.

Topic modelling has become a common approach to text analysis in the digital humanities (Blei, 2012). Our contribution is based on a project which aims to provide support for the analysis of a collection of textbooks used in schools in Germany in the 19th and early 20th century. The collection is comprised of 3,803 textbooks with 799,260 pages. Metadata is available over a wide range of attributes, including subject, type of school, date, and place of publication.

Prior user research (Heuwing, Mandl & Womser-Hacker, 2016) led to the conclusion that, in addition to an open-ended explorative analysis to find patterns and trends of interest, the most important task consists of a comparative analysis of the contents of the collection across sub-collections, either according to a dimension of metadata of interest, over time, or both. This kind of analysis is focused on a specific area of interest, which may be represented by several similar topics of the topic model. To this end two prototypical visualisations were created, one displaying topics and sub-collections as nodes in a network graph, the second one making use of heatmaps to compare correlations of topics in sub-collections. These tools were comparatively evaluated with researchers working in history with the aim to examine ways of analysing provided information (including metadata, topics and further metrics denoting their interdependencies, cf. sect. 4) for the collection when applying the visualisations to solve realistic tasks. Thereby, we try to shed some light on the following, explorative research question: *To which degree do both visualisation tools, also with respect to the represented data, support data analysis processes?*

2 Topic modelling

Topic Modelling describes a set of algorithms which help to analyse a large collection of documents based on its latent thematic structure. The most frequently used technique LDA (Latent Dirichlet Allocation) assumes that every document in the collection is generated from a fixed number of topics, each document exhibiting a different proportion of each topic. Every topic is defined as a distribution over all words within the document collection as a fixed vocabulary, giving high weights to those words that tend to co-occur (Blei, 2012). These properties of topic modelling enable the annotation of documents, while the aggregated distributions of topics are expected to resemble the thematic structure of the document collection. Based on that, exploration and further information foraging tasks over the document collection can be supported (ibid.). The potential for the field of digital humanities lies in the verification and formation of theories about a document collection of interest. However, a topic model generated on a collection should not be understood as an objective representation of its contents, but rather as a "lens for viewing a corpus of documents" (DiMaggio, Nag & Blei, 2013: 582) which is specific to the focus of the researcher and her research interests.

3 Visualisation of topic models

Interactive information visualisations are necessary when analysing complex data sets in an effective and efficient manner. The represented data can trigger insights that are relevant for a domain. Specifically, for the interpretation of topic models visualisations can provide an overview over the thematic structure of a collection and help to reveal relations between topics and between topics and sub-collections.

Only few previous approaches to support the analysis of sub-collections based on topic models have been described in the literature. *DiTop* (Oelke et al., 2014) is a tool that makes use of glyph based and spatial techniques to support the comparative analyses of up to three different sub-collections. In this context, glyphs distinguish *discriminative topics*, which are distinctive and characteristic for one of these sub-collections from *common topics*,

which are characteristic for all documents. These representations of topics are positioned according to their representativeness for the sub-collections.

Visualisations using network techniques are very popular for displaying topic models in general. They not only reflect relations between different entities in an intuitive manner, but they are also easily extensible to additional visual dimensions and thus suitable for representing multivariate data (Gretarsson et al., 2012). An example that includes documents and sub-collections of a collection is *TopicNets*. Collections of documents and topics are represented as nodes of the network, while the distribution of a topic in a collection is conveyed through the connecting edges. The similarity between topics is represented through node positioning, so that clusters of documents that exhibit similar topics can be identified (ibid.).

4 Prototyping visualisation techniques to compare topics

The implementation of a visualisation design is constrained by the underlying data set and its attributes (e.g. the dimensionality and type, either qualitative or quantitative) and should be optimised to support the most important user goals. For the project, the primary goal is to assist historians in proving or rejecting hypotheses about the contents of a collection of text documents in terms of relations between different sub-collections that are defined by attributes of the metadata of the documents. Activities of analysis that are necessary to reach this goal may include the analysis of relations between a selected set of topics relevant to the analysis and the analysis of differences in the relations of these topics to sub-collections. Accordingly, the analysis of topic distributions in sub-collections and their changes over time is a major concept. Based on the available data and the user goals that have been identified, for the comparative evaluation study two alternative visualisation designs have been implemented: Network graphs and small-multiples.

Figure 1 shows an example of a network graph employed in the study: Nodes of different types either depict topics, labelled with the three most common terms, or sub-collections according to a selected attribute of the documents. The different node types are assigned to specific colours. The overall intensity of a topic in the collection and the number of documents in a

sub-collection are mapped onto node size. Edges are connecting only nodes of different types, i.e. topics and sub-collections. The width of edges encodes the average topic intensity in a sub-collection, i.e. the average of the values of a topic in all documents of a sub-collection. Considering the layout, the proposition of Gretarsson et al. (2012) is adopted to preserve both information of relations between topics as well as between topics and the sub-collections of the corpus. First, the position of the topic nodes is fixed on the two-dimensional plane according to the similarity of topics based on multi-dimensional scaling, and second, a common graph layout algorithm is applied to position the nodes representing sub-collections within the topic space. According to this template, static graphs representing four different time intervals within the span of twenty years have been generated to analyse changes over time.

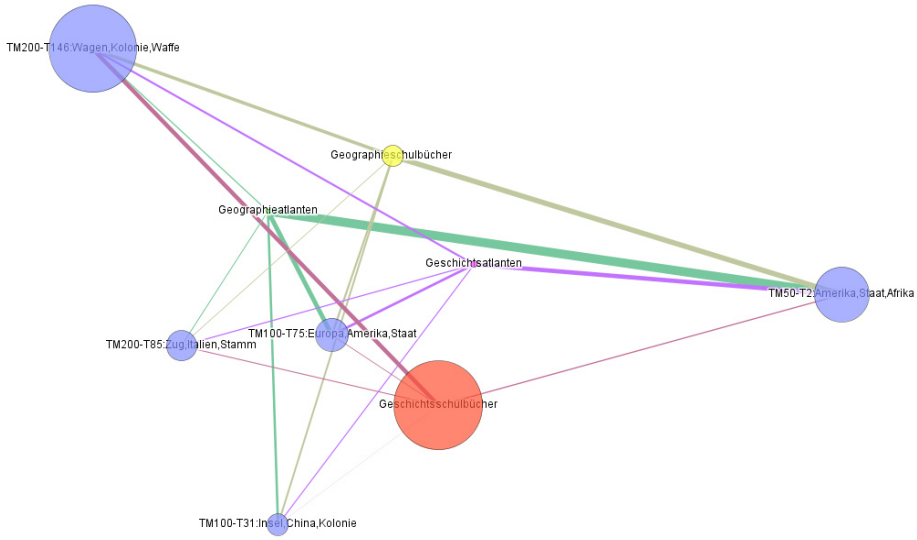

Fig. 1 Network graph consisting of topics and document sub-collections: blue nodes depict topics, other colours represent sub-collections (red for history textbooks, green for geographical atlases, purple for historical atlases, and yellow for geography textbooks)

The second design developed (cf. fig. 2) uses small-multiples to give a different perspective on the same data objects, by shifting the focus from relations between topics and document sub-collections towards relations between topics. Small-multiples are made up from small diagrams of the same

type and scale arranged in a grid according to different categories (Theus & Urbanek, 2008).

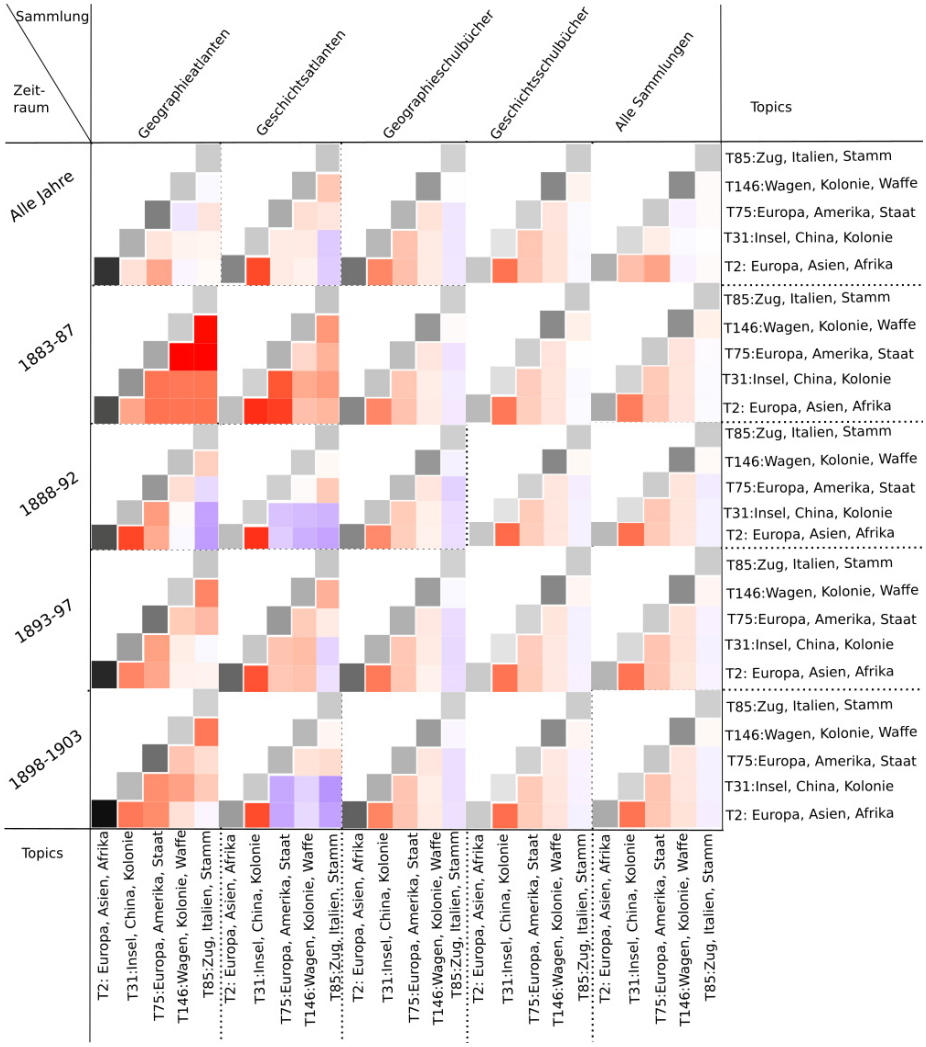

Fig. 2 Small-multiples consisting of heatmaps of topic correlations in sub-collections (columns: geographical atlases, historical atlases, geography textbooks, history textbooks, all collections; first row all years, followed by time intervals of five years)

In our visualisation, each consists of a matrix displaying the correlation of topics in the sub-collections. Pairwise correlation scores between topics were computed based on topic distributions over the documents for every interval in the time span of twenty years and for each sub-collection, as well as for the whole time span and over all sub-collections. Heatmaps were generated based on the resulting correlation matrices. As these are symmetrical, the upper half including the diagonal was left blank. Values from the respective correlation scores were mapped on an ordered, divergent colour map (ranging from blue for negative correlations over white for no correlations to red for positive correlations) to best represent the degree of correlation between two topics. The resulting heatmaps were then arranged in a grid according to their associated categories (time interval and document sub-collection). To maintain the information about the average topic intensity in a sub-collection, the rejected diagonal in the prior step was used by mapping this information onto an additional ordered, sequential colour map (using greyscales ranging from bright for very low topic intensity to dark for high topic intensity).

5 Methods

Task-based evaluation of visualisations tools can be used to study the fit between users' representation of the concepts and the representation chosen for the visualisations. Especially, when including both high-level tasks defined at a level of the users' goals in the domain and low-level tasks targeting the interaction with features of the visualisations, insights into the users' experience with the visualisations can be effectively gained (Faisal et al., 2007).

For the comparative evaluation of the two prototyped visualisation tools in the context of the above mentioned project, an empirical, task-based user study is conducted to capture user's experience from the target domain while interacting with the tools and the information conveyed by topics. Tasks are developed with analogies for both visualisations to be comparable. In the study, all users are presented with two low-level tasks for each tool to understand how they perceive visual elements and whether these are mapped on the intended data and attributes. General examples include "Which three topics are most likely commonly addressed over all sub-collections and which one distinguishes most from the rest?" and "Compare average topic intensi-

ties in the sub-collections history textbooks and geography textbooks." In addition, two high-level tasks for each tool address the characteristic of tasks in real conditions, e.g. being open-ended and focusing on the interpretation in the context of research questions, for example "How can the relations between topics during the time span of twenty years (1883–1903) be described? Based on that, what can be concluded on the development of the co-occurrence of topics addressing colonies and emigration over all sub-collections?" and "How can the relation of topic T85 to others at the course of the whole time span (1883–1903) be described between the sub-collections geography textbooks and history textbooks? Based on that, what can you conclude on the development of relations between topics addressing emigration and colonies within history textbooks compared to geography textbooks?" Task order was randomised with regard to the tools, with low-level tasks for each tool being presented before the high-level tasks.

Observations and the participants' explanations of their behaviour ("thinking aloud"-protocol) gave insights into the reasoning processes of participants during task completion. Results were selectively transcribed along with notes based on observations.

Five participants were selected in a high quality sample of expert users from the potential target groups of researchers in history. Each session took approximately an hour. At the beginning of each session, the rationale behind the study, topic modelling and the visualisations that were used were introduced. Every tool was provided as a static visualisation on the screen.

6 Results

Qualitative results were classified by making use of a systematic process of content analysis, thereby deriving a category system (Mayring, 2002). The following sections first present general activities of topic model interpretation, followed by the specific steps of interpreting the visual elements of the two visualisation tools together with common problems that occurred.

6.1 General activities when interpreting topic models

For the completion of tasks, participants applied different activities to understand the underlying models, i.e. topics in relation to other topics, their connections to sub-collections, and relating them to their own existing contextual knowledge. In the study, three participants derived new knowledge about the domain from the model based on these activities combined with prior conducted activities regarding the visual elements of the respective visualisations. For example, they draw a conclusion on the subjects addressed in a particular sub-collection, like one participant concluded on the content of history textbooks: "History textbooks differentiate between antique emigration and modern emigration along colonies."

Understanding relations between topics within sub-collections (2): The analysis of topics with similar or different distributions to each other was an important aspect for the interpretation of topics and their relations within sub-collections. A recognised similarity or dissimilarity between two topics within a sub-collection was explained by interpreting the content of topics using topic terms and then relating them to the features of the sub-collection. Hence, a low similarity of a topic to others within the sub-collection containing geography textbooks is explained by the fact that these are less likely to deal with socio-cultural topics, as one participant mentioned: "The topic [with the terms wagon, colony, weapon] does not play an important role within geography textbooks. […]. I assume that these address more topics similar to contents of geography than migrations, tribes and ethnicities." In one case, a participant recognised a high correlation between the topics with the terms *wagon*, *colony*, *weapon* and *migration*, *italy*, *tribe* in the sub-collection containing history textbooks, and concluded that history textbooks predominantly address the subjects of migration during the time of the Roman Empire rather than in the context of German colonies.

Contextual assignment of a topic to a sub-collection (2): Instead of taking only the represented data into account to relate a topic to a sub-collection, some participants referred to their own contextual knowledge. Consequently, a participant explains her own association of a topic to geography textbooks: "Regardless of the visualisation, I would connect [the topic containing the terms] europe, america, state more to geography textbooks than migration, italy, tribe. […] My association would be to connect geography textbooks in terms of the content more to states, than migration, italy, tribe. Tribe is something cultural and can't be classified into a geographical context."

Referring to topics: Topics were mostly referred to by the identifiers used in the visualisation. One participant explicitly named a topic antique topic, based on the interpretation of its terms, with the three most common *migration*, *italy*, *tribe*, as she explains her reasoning: "This [topic] addresses migration rather than emigration, the Italy campaign of Goths and Vandals […]. This is thousand years prior to these [topics]." In one case, the participant also named a topic *main geography topic*, based on its high intensity in the sub-collection of geography textbooks, thus transferring characteristics of the sub-collection to the topic.

6.2 Comparison of visualisations

The following categories describe activities conducted by the participants when interpreting the visual elements of the visualisations. Problems occurred especially at the beginning of tasks. During the test session, the participants' understanding of the visualisations and their confidence when interpreting them increased.

6.2.1 Network graphs

Process of gaining insight and generating information: Mostly, the visual elements, e.g. nodes or edges, were identified first. Then, further relevant attributes of these elements were interpreted, most importantly average topic intensity in a sub-collection conveyed through edge width, and topic similarity through position and proximities of topic nodes. Next, participants compared these attributes across time intervals. In one case, edge widths of all connections of a sub-collection node to all topic nodes were aggregated to describe the overall topic intensity in a sub-collection. These steps lead to the generation of information, like relations between topics or between sub-collections and topics (2), similarities or differences between attributes of topics or the average topic intensity (3), and their development over the time intervals (3).

Problems: Regarding the use of visual elements, problems were caused by the encoding of colour (e.g. assigning colour use to node types) (2), positions (e.g. meaning and arrangement of nodes in the plane was not always clear) (2) or the distinction between the encoding of topic similarity and average topic intensity (1). In the high level tasks, participants often created mental models that did not always match the intended meaning of visual elements.

For example, centrality of topic nodes was interpreted as their overall relevance (1), while closeness was used to describe the relations between topics and sub-collections instead of using edge widths (2). Position was also applied to describe changes over time and the development of relations (2). Edges and their widths were not only used to describe the relation of topics and sub-collections, but also the relation between topics. In one case, a topic node with comparatively narrow edges to all of the sub-collection nodes was interpreted as not being relevant in the context of any other topic.

6.2.2 Small-multiples

Process of gaining insight and generating information: The initiation of an analysis usually consisted of identifying the colour map, either for topic correlations or average topic intensity (2). Next, participants tried to identify relevant categories. They either considered all categories in each time interval and sub-collection or used only the overview over the complete time span or all sub-collections. Relevant correlations between topics within specific categories were first identified in general (4). Similar to the use of network graphs, results of average topic intensities were sometimes aggregated, e.g. for a sub-collection (2). Correlations of several identified categories were compared next. A sub-activity here was to understand the impact of a correlation between two topics in a particular sub-collection on those in all sub-collections. Comparing different categories that way also resulted in finding interesting patterns (similarities or differences) regarding relations between topics and categories (either sub-collections or time interval or both) or correlations between topics (5). Investigating these patterns particularly resulted in recognising changes over time (2).

Problems: Commonly faced problems included matching the right topic pairs to each cell in the heatmap (3), recognising differences in saturation and luminance of the colour maps for precisely discriminating correlation scores (4) and creating an understanding for the generation of heatmaps by the underlying computation of the correlation scores (2).

6.3 Summary

Processes of analysis appear to be influenced by the form of visualisations. When compared to network graphs, small-multiples enable finer grained approaches for analysis and mainly support the generation of domain related

knowledge through the representation of correlations between a pair of topics in a particular sub-collection. Network graphs, on the other hand, tend to support more insights and provide a comprehensive overview of the relations of topics and sub-collections. Participants seem to readily interpret every visual feature to derive conclusions about relationships in the data, especially if their findings are congruent with existing background knowledge. But using complex graphs that include many visual dimensions also entails risks because elements of the visual syntax are more likely to be misinterpreted.

7 Conclusion

The participation of representatives of the very specialised domain of historical textbook researchers together with the qualitative analysis of results provide insights into the mental processes applied during the analysis which are highly relevant for the project. As the results show, historians reveal many approaches of analysing information about topic models, where topic terms play a major role. Activities in this context are also mainly responsible for deriving general conclusions on the content of particular sub-collections compared to others and thus creating domain related knowledge.

The analysis of the activities for the interpretation of the models and visualisations highlight opportunities for improving our visualisations of topic models. For example, the strategies to compare distributions and correlations of topics within sub-collections to the respective values of the complete collection indicate the need for comparative metrics (e.g. higher/lower than general), while the important role of external, contextual knowledge indicates the need to annotate elements with the users' own terminology. Qualitative findings reveal that network graphs are less difficult to understand but may also lead to mis- and over-interpretations of the data provided by topic models. This finding is probably also relevant for other areas where visualisations of topic models are used as a basis for further interpretations.

Acknowledgment

We would like to thank all the participants of the study for their support and input. This research was partly funded by a SAW grant from the Leibniz Association.

References

Blei, D. M. (2012): Probabilistic Topic Models. In: *Communications of the ACM*, 55 (4). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/download/4645</5021

DiMaggio, P., M. Nag, and D. M. Blei (2013): Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. In: *Poetics*, 41 (6), 570–606. http://linkinghub.elsevier.com/retrieve/pii/S0304422X13000661

Faisal, S., P. Cairns, and A. Blandford (2007): Challenges of Evaluating the Information Visualisation Experience. In: *BCS-HCI '07, Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI ... but not as we know it – Volume 2* (pp. 167–170). University of Lancaster: BCS Learning & Development Ltd. Swindon. http://dl.acm.org/citation.cfm?id=1531407.1531451

Gretarsson, B., J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman et al. (2012): TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. In: *ACM Trans. Intell. Syst. Technol.*, 3 (2). http://doi.acm.org/10.1145/2089094.2089099

Heuwing, B., T. Mandl, and C. Womser-Hacker (2016): Combining contextual interviews and participative design to define requirements for text analysis of historical media. In: *ISIC: The Information Behaviour Conference*. Zadar

Mayring, P. (2002): *Einführung in die qualitative Sozialforschung: eine Anleitung zu qualitativem Denken* (5th ed.). Weinheim and Basel: Beltz.

Oelke, D., H. Strobelt, C. Rohrdantz, I. Gurevych, I., and O. Deussen (2014): Comparative Exploration of Document Collections: A Visual Analytics Approach. In: *Computer Graphics Forum*, 33 (3), 201–210. http://dx.doi.org/10.1111/cgf.12376

Theus, M. and S. Urbanek (2008): *Interactive Graphics for Data Analysis: Principles and Examples*. Boca Raton, FL: Chapman & Hall/CRC.

Zugal, S., J. Pinggera, H. Reijers, M. Reichert, and B. Weber (2012): Making the Case for Measuring Mental Effort. In: Chaudron et al. (Eds.): *EESSMod '12, Proceedings of the Second Edition of the International Workshop on Experiences and Empirical Studies in Software Modelling* (pp. 37–43). New York: ACM. http://doi.acm.org/10.1145/2424563.2424571